Clinical Natural Language Processing in languages other than English: opportunities and challenges

Névéol *et al. Journal of Biomedical Semantics* (2018) 9:12
https://doi.org/10.1186/s13326-018-0179-8

Journal of Biomedical Semantics

REVIEW Open Access

Clinical Natural Language Processing in languages other than English: opportunities and challenges

Aurélie Névéol[1]*, Hercules Dalianis[2], Sumithra Velupillai[3,4], Guergana Savova[5] and Pierre Zweigenbaum[1]

Abstract

Background: Natural language processing applied to clinical text or aimed at a clinical outcome has been thriving in recent years. This paper offers the first broad overview of clinical Natural Language Processing (NLP) for languages other than English. Recent studies are summarized to offer insights and outline opportunities in this area.

Main Body: We envision three groups of intended readers: (1) NLP researchers leveraging experience gained in other languages, (2) NLP researchers faced with establishing clinical text processing in a language other than English, and (3) clinical informatics researchers and practitioners looking for resources in their languages in order to apply NLP techniques and tools to clinical practice and/or investigation. We review work in clinical NLP in languages other than English. We classify these studies into three groups: (i) studies describing the development of new NLP systems or components de novo, (ii) studies describing the adaptation of NLP architectures developed for English to another language, and (iii) studies focusing on a particular clinical application.

Conclusion: We show the advantages and drawbacks of each method, and highlight the appropriate application context. Finally, we identify major challenges and opportunities that will affect the impact of NLP on clinical practice and public health studies in a context that encompasses English as well as other languages.

Keywords: Natural Language Processing, Clinical Decision-Making, Languages other than English

Background

Clinical research in a global context

Healthcare is a top priority for every country. The goal of clinical research is to address diseases with efforts matching the relative burden [1]. Computational methods enable clinical research and have shown great success in advancing clinical research in areas such as drug repositioning [2]. Much clinica l information is currently contained in the free text of scientific publications and clinical records. For this reason, Natural Language Processing (NLP) has been increasingly impacting biomedical research [3–5]. Prime clinical applications for NLP include assisting healthcare professionals with retrospective studies and clinical decision making [6, 7]. There have been a number of success stories in various biomedical NLP applications in English [8–19]. The ability to analyze clinical text in languages other than English opens access to important medical data concerning cohorts of patients who are treated in countries where English is not the official language, or in generating global cohorts especially for rare diseases. One such example is the Phelan-McDermid Syndrome Foundation (PMSF), which is leading a Patient Powered Research Network project (part of the Patient Centered Outcome Research Institute, PCORI [20] on a very rare disease. PMSF parents, together with researchers and advisors, launched an international patient registry, the PMSIR, that is directed, governed, and implemented by patient families. There are a total of 900 cases of this rare disease in the entire world. Each patient contributed their EHR and genomics data to enable phenotype/genotype studies. Recently, Kohane et al. have shown that methods allowing an aggregated

*Correspondence: aurelie.neveol@limsi.fr
[1]LIMSI, CNRS, Université Paris Saclay, Rue John von Neumann, F-91405 Orsay Paris, France
Full list of author information is available at the end of the article

exploitation of clinical data from multiple healthcare centers could contribute to make headway in the understanding of autism spectrum disorders [21]. Cross-lingual text mining of newswires in thirteen languages was shown to be helpful for automated health surveillance of disease outbreaks, and was routinely implemented in the BioCaster portal [22].

In this context, data extracted from clinical text and clinically relevant texts in languages other than English adds another dimension to data aggregation. The World Health Organization (WHO) is taking advantage of this opportunity with the development of IRIS [23], a free software tool for interactively coding causes of death from clinical documents in seven languages. The system comprises language-dependent modules for processing death certificates in each of the supported languages. The result of language processing is standardized coding of causes of death in the form of ICD10 codes, independent of the languages and countries of origin.

Objective and Scope

This paper follows-up on a panel discussion at the 2014 American Medical Informatics Association (AMIA) Fall Symposium [24]. Following the definition of the International Medical Informatics Association (IMIA) Yearbook [25, 26], clinical NLP is a sub-field of NLP applied to clinical texts or aimed at a clinical outcome. This encompasses NLP applied to texts in Electronic Health Records (EHRs), but also extends to the development of resources for clinical NLP systems, and to clinically relevant research addressing biomedical information retrieval or the analysis of patient-authored text for public health or diagnostic purposes. We survey studies conducted over the past decade and seek to provide insight on the major developments in the clinical NLP field for languages other than English. We outline efforts describing (i) building new NLP systems or components from scratch, (ii) adapting NLP architectures developed for English to another language, and (iii) applying NLP approaches to clinical use cases in a language other than English.

Finally, we identify major NLP challenges and opportunities with impact on clinical practice and public health studies accounting for language diversity.

Main Text

Review method and selection criteria

Conducting a comprehensive survey of clinical NLP work for languages other than English is not a straightforward task because relevant studies are scattered across the literature of multiple fields, including medical informatics, NLP and computer science. In addition, the language addressed in these studies is not always listed in the title or abstract of articles, making it difficult to build search queries with high sensitivity and specificity.

In order to approximate the publication trends in the field, we used very broad queries. A Pubmed query for "Natural Language Processing" returns 4,486 results (as of January 13, 2017). Table 1 shows an overview of clinical NLP publications on languages other than English, which amount to almost 10% of the total.

We are showing the results of this query as an imperfect proxy for estimating the scale of the biomedical literature relevant to NLP research, as some publications addressing clinical NLP may not appear in PubMed, and some publications referenced in PubMed may be missed by the query. As described below, our selection of studies reviewed herein extends to articles not retrieved by the query.

Figure 1 shows the evolution of the number of NLP publications in PubMed for the top five languages other than English over the past decade. We can see that French benefits from a historical but sustained and steady interest. Chinese and Spanish have recently attracted sustained efforts. Japanese and German seem to receive plateauing attention.

This work is not a systematic review of the clinical NLP literature, but rather aims at presenting a selection of studies covering a representative (albeit not exhaustive) number of languages, topics and methods. We browsed the results of broad queries for clinical NLP in MEDLINE and ACL anthology [26], as well as the table of contents of the recent issues of key journals. We also leveraged our own knowledge of the literature in clinical NLP in languages other than English. Finally, we solicited additional references from colleagues currently working in the field.

Our selection criteria were based on the IMIA definition of clinical NLP [25, 26]. For instance, the broad queries employed in MEDLINE resulted in a number of publications reporting work on speech or neurobiology, not on clinical text processing, which we excluded. Moreover, with the increased volume of publications in this area in the last decade, we prioritized the inclusion of studies from the past decade. In total, 114 publications across a wide range of languages fulfilled these criteria (Table 1).

Clinical NLP in languages other than English

This section reviews the topics covered by recently published research on clinical NLP which addresses languages other than English. We organize the section by the type of strategies used in the specific studies. Table 2 presents a classification of the studies cross-referenced by NLP method and language.

Building new systems and resources

New NLP systems or components Some of the work in languages other than English addresses core NLP tasks that have been widely studied for English, such as sentence boundary detection [27], part of speech tagging [28–30], parsing [31, 32], or sequence segmentation [30].

Table 1 Number of publications returned by a PubMed search for "Natural Language Processing AND *language* [tiab]" where *language* is instantiated with a specific language name, on January 13, 2017 along with references cited in this review for each language. The last row (bolded) presents overall information for all languages studied in this review

Language (ISO 639-1 language code)	PubMed Count	Cited in this review
French (FR)	111	[31, 77]*[71, 160, 161]*[158]*[65]*[78]*[66, 79, 94] [156]* [50, 67, 89, 109, 120]* [54]* [154]* [159]* [140] [138]* [7, 56, 59, 60, 90, 116, 117] [163]* [112] [70]* [126, 152, 153] [55]*
German (DE)	69	[31]*[115] [72]*[156]* [154]*[141] [109]*[118] [138]* [84] [163]* [27, 53, 88] [70]* [80, 124] [36, 106]
Chinese (ZH)	54	[155]* [68, 73, 96] [154]* [42, 43, 69, 99] [103, 122]
Spanish (ES)	39	[161]*[158]*[155]*[156]*[54]*[154]*[138] *[30, 98, 107, 119] [70]*[58] [34, 108], [55, 157]*
Japanese (JA)	30	[158]*[33, 37] [154]* [49, 127, 149, 151],
Dutch (DU)	20	[114] [139] [138]*[110] [70]*
Swedish (SV)	15	[57, 104] [74]*[92] [109]* [48, 61, 105] [35, 93, 113, 123]
Portuguese (PT)	14	[28, 83], [55]*
Greek (EL)	14	[52]
Italian (IT)	12	[46, 47, 97]
Korean (KO)	11	[155]*[91]
Arabic (AR)	9	[158]*[162]
Finnish (FI)	9	[38, 40] [74]*[32, 85, 121]
Czech (CS), Russian (RU)	7	[155]*, [163]*
Polish (PL)	6	[29, 82], [156]*
Hebrew (HE)	5	[41, 44]
Danish (DA)	4	[86, 87] [45]
Turkish (TR)	3	[156]*
Bulgarian (BG)	2	[62, 64, 95, 100–102]
Basque (EU)	1	[51]
Georgian (KA)	1	[125]
Hungarian (HU)	0	[156]*
Overall	**435**	**114**

Note that some included articles are not indexed in MEDLINE but in other publication venues such as ACL. A star indicates work that addresses several languages

Word segmentation issues are more obviously visible in languages which do not mark word boundaries with clear separators such as white spaces. This is the case, for instance, in Chinese, Japanese, Vietnamese and Thai. A study of automatic word segmentation in Japanese addressed the lack of spacing between words in this language [33]. The authors implemented a probabilistic model of word segmentation using dictionaries. Abbreviations are common in clinical text in many languages and require term identification and normalization strategies. These have been studied for Spanish [34], Swedish [35], German [27, 36] and Japanese [37]. More complex semantic parsing tasks have been addressed in Finnish [38] through the addition of a PropBank layer [39] to clinical Finnish text parsed by a dependency parser [40].

Core NLP tasks are sometimes evaluated as part of more complex tasks. For instance, a study on Hebrew medical text shows that segmentation methods accounting for transliterated words yield up to 29% performance improvement in medical term extraction [41]. Word segmentation was also shown to outperform character segmentation for named entity recognition in Chinese clinical text. In addition, performing segmentation and named entity recognition jointly yielded a 1%

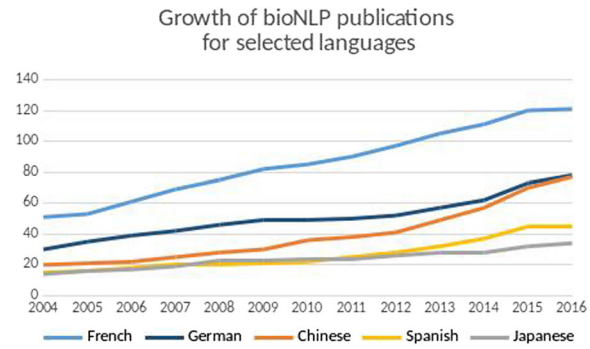

Fig. 1 Growth of bio-clinical NLP publications in MEDLINE over the past decade, for the top 5 studied languages other than English

improvement for both. The overall performance of named entity recognition using these special features was above 0.90 F1-measure for four entity types, a performance comparable to English state-of-the-art [42, 43]. Conversely, in an effort addressing the expansion of English abbreviations in Japanese text [37] a study on eight short forms associated to two or more long forms found that character (vs. word) segmentation performed better for the task. However, it can be argued that in the context of code-switching and transliteration (English abbreviations appeared verbatim in Japanese text, accompanied by an expanded form of the acronym in Japanese), the distribution of words and characters made the text sufficiently different from standard Japanese to warrant specific processing. Cohen et al. [44] studied the impact of the high frequency of transliterated terms in Hebrew clinical narratives. They report that the use of a semi-automatically acquired medical dictionary of transliterated terms improves the performance of information extraction. The effect of spelling correction and negation detection on an ICD10 coding system was studied for Danish and both features were found to yield improved performance [45].

Lexicons, terminologies and annotated corpora While the lack of language specific resources is sometimes addressed by investigating unsupervised methods [46, 47], many clinical NLP methods rely on language-specific resources. As a result, the creation of resources such as synonym or abbreviation lexicons [27, 36, 48] receives a lot of effort, as it serves as the basis for more advanced NLP and text mining work.

Distributional semantics was used to create a semantic space of Japanese patient blogs, seed terms from the categories Medical Finding, Pharmaceutical Drug and Body Part were used to expand the vocabularies with promising results [49].

There is sustained interest in terminology development and the integration of terminologies and ontologies in the UMLS [50], or SNOMED-CT for languages such as Basque [51]. In other cases, full resource suites including terminologies, NLP modules, and corpora have been developed, such as for Greek [52] and German [53].

The development of reference corpora is also key for both method development and evaluation. Recently, researchers produced annotated corpora for tasks such as machine translation [54, 55], de-identification in French [56] and Swedish [57], drug-drug interaction in Spanish [58], named entity recognition and normalization for French [59], and also for linguistic elements such as verbal propositions and arguments for Finnish [38]. The study of annotation methods and optimal uses of annotated corpora has been growing increasingly with the growth of statistical NLP methods [7, 60, 61].

For some languages, a mixture of Latin and English terminology in addition to the local language is routinely used in clinical practice. This adds a layer of complexity to the task of building resources and exploiting them for downstream applications such as information extraction. For instance, in Bulgarian EHRs medical terminology appears in Cyrillic (Bulgarian terms) and Latin (Latin and English terms). This situation calls for the development of specific resources including corpora annotated for abbreviations and translations of terms in Latin-Bulgarian-English [62]. The use of terminology originating from Latin and Greek can also influence the local language use in clinical text, such as affix patterns [63].

Multilingual corpora are used for terminological resource construction [64] with parallel [65–67] or comparable [68, 69] corpora, as a contribution to bridging the gap between the scope of resources available in English vs. other languages. More generally, parallel corpora also make possible the transfer of annotations from English to other languages, with applications for terminology development as well as clinical named entity recognition and normalization [70]. They can also be used for comparative evaluation of methods in different languages [71].

A notable use of multilingual corpora is the study of clinical, cultural and linguistic differences across countries. A study of forum corpora showed that breast cancer information supplied to patients differs in Germany vs. the United Kingdom [72]. Furthermore, a study of clinical documents in English and Chinese evidenced a lower density of treatment concepts in Chinese documents [73] which was interpreted as a reflection of cultural differences between clinical narrative styles and suggests that this needs to be accounted for when designing clinical NLP systems for Chinese.

Conversely, a comparative study of intensive care nursing notes in Finnish vs. Swedish hospitals showed that

Table 2 List of studies presented in this review categorized by NLP method used and language(s) addressed

Method/Task	Language/reference cited in this review
Core NLP	
- Morphology	FR [78] PL [29]
- Part of Speech tagging	PT [28] ES [30]
- Parsing	FI [32, 38, 40] FR [31, 77] GR [52] JA [33, 37]
- Segmentation	DE [27] HE [41]
Resource development	
- Lexicons	BG [62] EL [52] EU [51] FR [50, 65–67] HE [44] JA [49] SV[48] ZH [68, 69]
- Corpora and annotation	EL [52] EN-{FR,ES} [54] EN-{ES,FR,PT} [55] ES [58] FR [59, 117]
- Models, methods	DE [53] FR [60]
De-identification	FR [56, 79, 89, 90] KO [91] SV [57, 92]
Information extraction	
- Medical Concepts	BG [62, 64] ZH [42, 43] DE [80, 84, 115, 124] DU [114] ES [34] IT [46, 47, 97] PL [82] SV [61]
- Findings/Symptoms	DE [118], SV [61, 93] ZH [96]
- Drugs/Adverse events	BG [95, 102] DA [87] ES [98] FR [94, 116] SV [61]
- Specific characteristics	EN-{ZH,FR,DE,JA,ES} [154] FR [120] ZH [99] DU [114]
- Relations	BG [64] DE [84, 115] IT [46, 47]
Classification	
- Phenotyping from EHR text	BG [100] ES [119] FI [121] FR [7, 126] KA [125] PT [83] SV [123] ZH [122]
- Indexing and coding	EN-FR [71] FI [85] FR [153] JA [149, 151]
- Patient-authored text	JA [127]
- Cohort stratification	DA [86] DE [88]
Context Analysis	DU [110]
- Negation detection	BG [101] DA [45] DE [106] DU [110] ES [107, 108] FR,DE,SV [109] SV [104, 105]
- Uncertainty/Assertion	SV [105] ZH [103]
- Temporality	FR [112] SV [113]
- Abbreviation	DE [36] SV [35]
- Experiencer	DU [110]
Multilingual tasks	
- Translation	EN-ES [157] EN-FR [159] EN-{KO,RU,ES,ZH} [155] EN-{FR,DE,HU,PL,ES,TU} [156], FR-DE [158]
- Information Retrieval	AR [162] FR [160], EN-{CZ,DE,FR} [163] EN-{ES,FR} [161]
- Cultural analysis	DE [72], EN-ZH [73], FI-SV [74]
Shared tasks	
- CLEF-ER2013	DE,DU,FR,ES- [138]
- CLEF eHealth 2015, 2016	FR [152, 153]
- NTCIR 2014, 2016	JA [149, 151]

The two letter language codes are introduced in Table 1. When multiples languages are addressed in one paper we provide a comma separated list; dashes mark language pairs in multilingual work

there are essentially linguistic differences while the content and style of the documents is similar [74].

Adapting NLP architectures developed for English

Studying sublanguages, Harris [75] observed that "The structure of each science language is found to conform to the information in that science rather than to the grammar of the whole language." Sager's LSP system [76], developed for the syntactic analysis of medical English, was adapted to French [77]. Deléger et al. [78] also describe how a knowledge-based morphosemantic parser could be ported from French to English.

This shows that adapting systems that work well for English to another language could be a promising path. In practice, it has been carried out with varying levels of success depending on the task, language and system design. The importance of system design was evidenced in a study attempting to adapt a rule-based de-identification method for clinical narratives in English to French [79]. Language-specific rules were encoded together with de-identification rules. As a result, separating language-specific rules and task-specific rules amounted to re-designing an entirely new system for the new language. This experience suggests that a system that is designed to be as modular as possible, may be more easily adapted to new languages. As a modular system, cTAKES raises interest for adaptation to languages other than English. Initial experiments in Spanish for sentence boundary detection, part-of-speech tagging and chunking yielded promising results [30]. Some recent work combining machine translation and language-specific UMLS resources to use cTAKES for clinical concept extraction from German clinical narrative showed moderate performance [80]. More generally, the use of word clusters as features for machine learning has been proven robust for a number of languages across families [81].

Similarly to work in English, the methods for Named Entity Recognition (NER) and Information Extraction for other languages are rule-based [82, 83], statistical, or a combination of both [84]. With access to large datasets, studies using unsupervised learning methods can be performed irrespective of language, as in Moen et al. [85] where such methods were applied for information retrieval of care episodes in Finnish clinical text. Knowledge-based methods can be applied when terminologies are available, e.g. extending information contained in structured data fields with information from Danish clinical free-text with dictionary-based approaches for the study of disease correlations [86] or adverse events [87]. For German, extracting information from clinical narratives for cohort building using simple rules was successful [88].

NER essentially focuses on two types of entities: personal health identifiers in the context of clinical document de-identification [56, 57, 79, 79, 89–92] and clinical entities such as diseases, signs/symptoms [93], procedures or medications [61, 94–100], as well as their context of occurrence: negation [101], assertions [102, 103] and experiencer (i.e. whether the entities are relevant to the patient or a third party such as a family member or organ donor).

Systems addressing a task such as negation may be easily adapted between languages of the same family that express negation using similar syntactic structures as is the case for English and Swedish [104, 105], English and German [106], English and Spanish [107, 108], or even English, French, German and Swedish [109]. However, it can be difficult to pinpoint the reason for differences in success for similar approaches in seemingly close languages such as English and Dutch [110].

Another important contextual property of clinical text is temporality. Heideltime is a rule-based system developed for multiple languages to extract time expressions [111]. It has been adapted for clinical text in French [112] and Swedish [113].

Global concept extraction systems for languages other than English are currently still in the making (e.g. for Dutch [114], German [115] or French [116, 117]).

The entities extracted can then be used for inferring information at the sentence level [118] or record level, such as smoking status [119], thromboembolic disease status [7], thromboembolic risk [120], patient acuity [121], diabetes status [100], and cardiovascular risk [122].

Applications

There are a number of studies describing applications relying on some NLP preprocessing. Jacobson et al. [123] use deep learning to detect healthcare associated infections in Swedish patient records. Lopprich et al [124] describe a system using NLP methods for German to classify the diagnoses of Multiple Myeloma patients at Heidelberg University Hospital. The high average F1-scores demonstrate the suitability of the investigated methods. However, it was also shown that there is no best practice for an automatic classification of data elements from free-text diagnostic reports. A study on Georgian medical records, where documents were classified into types (Ultrasonography, X-ray and Endoscopy) and clinical categories (e.g. Thyroid, Biliary system) showed promising results, and highlights early work in an understudied, highly agglutinative language [125].

Metzger et al. [126] show how the development of machine learning-based classifiers using free-text data can be used to identify suicide attempts in a French Emergency Department with promising results (70.4-95.3% F1), demonstrating that the quality of epidemiological

indicators can be improved by these types of approaches as opposed to manually coded information. Grouin et al. [120] show that information extraction from clinical records can sucessfully be used to automatically compute a cardio-vascular alert score on par with experts. Similarly, Takano et al. [127] use NLP to analyze Japanese patients cue-recalled memories to automatically determine memory specificity, an important indicator in the diagnosis of memory dysfunctions. NLP-based systems have been integrated into a clinical workflow for assisting clinical decision making or contributing to the construction of large health information system such as data warehouses. For instance, the Bulgarian system *BITool* is used for the construction of the register of diabetic patients in Bulgaria, which contains over 100 million de-identified reimbursement requests from all general practitioners and specialists in the country for a 3 year period [100].

Discussion

As we enter an era where big data is pervasive and EHRs are adopted in many countries, there is an opportunity for clinical NLP to thrive beyond English, serving a global role.

How to develop a clinical NLP application in a language other than English?

Research on the use of NLP for targeted information extraction from, and document classification of, EHR text shows that some degree of success can be achieved with basic text processing techniques. It can be argued that a very shallow method such as lexicon matching/regular expressions to a customized lexicon/terminology is sufficient for some applications [128]. For tasks where a clean separation of the language-dependent features is possible, porting systems from English to structurally close languages can be fairly straightforward. On the other hand, for more complex tasks that rely on a deeper linguistic analysis of text, adaptation is more difficult.

In summary, the level of difficulty to build a clinical NLP application depends on various factors including the difficulty of the task itself and constraints linked to software design. Legacy systems can be difficult to adapt if they were not originally designed with a multi-language purpose.

Where are the best opportunities?

Clinical NLP in any language relies on methods and resources available for general NLP in that language, as well as resources that are specific to the biomedical or clinical domain.

In this respect, English is by far the most resource-rich language, with advanced tools dedicated to the biomedical domain such as part-of-speech taggers (e.g. MedPOST [129]), parsers (e.g. GATE [130], Charniak-McClosky [131], enju [132]), biomedical concept extractors (e.g. MetaMap [133], cTAKES [134, 135], NCBO [136]). For other languages, data and resources are sometimes scarce.

The UMLS (Unified Medical Language System [137]) aggregates more than 100 biomedical terminologies and ontologies. In its 2016AA release, the UMLS Metathesaurus comprises 9.1 million terms in English followed by 1.3 million terms in Spanish. For all other languages, such as Japanese, Dutch or French, the number of terms amounts to less than 5% of what is available for English. Additional resources may be available for these languages outside the UMLS distribution. Details on terminology resources for some European languages were presented at the CLEF-ER evaluation lab in 2013 [138] for Dutch [139], French [140] and German [141].

Medical ethics, translated into privacy rules and regulations, restrict the access to and sharing of clinical corpora. Some datasets of biomedical documents annotated with entities of clinical interest may be useful for clinical NLP [59]. However, there are currently no sharable clinical datasets comparable to the i2b2 datasets [142, 143], the ShARe corpus [144], the THYME corpus [145, 146] or the MIMIC corpus [147] in languages other than English except the Turku Clinical TreeBank and PropBank [32, 38, 148] in Finnish and the small subset of 100 patient pseudonymized records in the Stockholm EPR PHI Pseudo Corpus [92] in Swedish, and the examinations clinical texts of the MedNLPDoc corpus in Japanese [149], albeit only with document-level annotation.

Past experience with shared tasks in English has shown international community efforts were a useful and efficient channel to benchmark and improve the state-of-the-art [150]. The NTCIR-11 MedNLP-2 [151] and NTCIR-12 MedNLPDoc [149] tasks focused on information extraction from Japanese clinical narratives to extract disease names and assign ICD10 codes to a given medical record. The CLEF-ER 2013 evaluation lab [138] was the first multi-lingual forum to offer a shared task across languages. It resulted in a small multi-lingual manually-validated reference dataset [70] and prompted the development of a large gold-standard annotated corpus of clinical entities for French [59], currently in use in a clinical named entity recognition and normalization task in the CLEF eHealth evaluation lab [152, 153]. Our hope is that this effort will be the first in a series of clinical NLP shared tasks involving languages other than English. The establishment of the health NLP Center as a data repository for health-related language resources (www.center.healthnlp.org) will enable such efforts.

In summary, there is a sharp difference in the availability of language resources for English on one hand, and other languages on the other hand. Corpus and terminology development are a key area of research for languages

other than English as these resources are crucial to make headway in clinical NLP.

How do we best leverage existing data and tasks?

Leveraging resources for English. The resource availability for English has prompted the use of machine translation as a way to address resource sparsity in other languages. Off-the-shelf automatic translators, e.g. Google translate, were found to have the potential to reduce language bias in the preparation of randomized clinical trials reports language pairs [154]. However, it was shown to be of little help to render medical record content more comprehensible to patients [155]. A systematic evaluation of machine translation tools showed that off-the-shelf tools were outperformed by customized systems [156]; however, this was not confirmed when using a smaller in-domain corpus [157]. Encouragingly, medical speech translation was shown to be feasible in a real clinical setting, if the system focused on narrowly-defined patient-clinician interactions [158]. Further work focused on acquiring and evaluating targeted resources [54, 55, 159].

Machine translation is used for cross-lingual Information Retrieval to improve access to clinical data for non-native English speakers. Successful query translation (for a limited set of query terms) was achieved for French using a knowledge-based method [160]. Query translation relying on statistical machine translation was also shown to be useful for information retrieval through MEDLINE for queries in French, Spanish [161] or Arabic [162]. More recently, custom statistical machine translation of queries was shown to outperform off-the-shelf translation tools using queries in French, Czech and German on the CLEF eHealth 2013 dataset [163]. Interestingly, while the overall cross-lingual retrieval performance was satisfactory, the authors found that better query translation did not necessarily yield improved retrieval performance.

More recently, machine translation was also attempted to adapt and evaluate cTAKES concept extraction to German [80], with very moderate success. Making use of multilingual resources for analysing a specific language seems to be a more fruitful approach [152, 153, 164]. It also yielded improved performance for word sense disambiguation in English [165].

Learning from other languages. The common clinical NLP research topics across languages prompt a reflexion on clinical NLP in a more global context.

Recent work on negation detection in English clinical text [166] suggests that the ability to successfully address a particular clinical NLP task on a particular corpus does not necessarily imply that the results can be generalized without significant adaptation effort. This may hold true for adaptations across languages as well, and suggests a direction for future work in the study of language-adaptive, domain-adaptive and task-adaptive methods for clinical NLP. The LORELEI [167] initiative aims to create NLP technologies for languages with low resources. While not specific to the clinical domain, this work may create useful resources for clinical NLP.

Interestingly, segmentation with lack of spacing in Japanese [33] could be successfully applied to English text where spacing between words was removed such as in Character Recognition (OCR) where word spacing is often not captured properly. Duque et al. [165] show that multilingual ressources can be useful for processing English text: for a word sense disambiguation task, multilingual resources yield a 7% improvement in performance, compared to monolingual resources.

Conclusion

In summary, we find a steady interest in clinical NLP for a large spectrum of languages other than English that cover Indo-European languages such as French, Swedish or Dutch as well as Sino-Tibetan (Chinese), Semitic (Hebrew) or Altaic (Japanese, Korean) languages. Our review of recent studies shows that (1) the field is maturing, (2) researchers in the community have access to datasets, which enables them to develop powerful methods to address clinical NLP tasks of interest such as EHR de-identification, clinical entity recognition, normalization and contextualization. We identified the need for shared tasks and datasets enabling the comparison of approaches within- and across- languages. Furthermore, the challenges in systematically identifying relevant literature for a comprehensive survey of this field lead us to also encourage more structured publication guidelines that incorporate information about language and task. We suggest that efforts in analyzing the specificity of languages and tasks could contribute to methodological advances in adaptive methods for clinical NLP.

Abbreviations

ACL: Association for computational linguistics; AMIA: American medical informatics association; CLEF: Cross language evaluation forum; cTAKES: clinical Text analysis and knowledge extraction system; EHR: Electronic health record; ICD10: International Classification of Diseases, 10th revision; IMIA: International medical informatics association; NLP: Natural language processing; SNOMED-CT: Systematized nomenclature of medicine - clinical terms; UMLS: Unified medical language system; WHO: World health organization

Acknowledgments

The authors would like to thank Galja Angelova and Svetla Boycheva for their knowledgeable insight on clinical NLP work on Bulgarian.

Funding

This work was supported by the French National Agency for Research under grant CABeRneT[1] ANR-13-JS02-0009-01; US National Institutes of Health funding (U24CA184407, LM 10090, GM114355); the Swedish Research Council (2015-00359) and the Marie Skłodowska Curie Actions, Cofund, Project INCA 600398.

Availability of data and materials

Not applicable.

Authors' contributions
AN wrote a first draft of the manuscript. All authors sought relevant references to be added and each contributed to the creation of Table 2. All authors contributed to the writing process and approved the final version of the manuscript.

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]LIMSI, CNRS, Université Paris Saclay, Rue John von Neumann, F-91405 Orsay Paris, France. [2]DSV, Stockholm University, Kista, Sweden. [3]School of Computer Science and Communication, KTH, Stockholm, Sweden. [4]Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK. [5]Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts, USA.

Received: 22 May 2017 Accepted: 14 February 2018
Published online: 30 March 2018

References

1. Emdin C, Odutayo A, Hsiao A, Shakir M, Hopewell S, Rahimi K, Altman D. Association between randomised trial evidence and global burden of disease: cross sectional study (Epidemiological Study Of Randomized Trials–ESORT). BMJ. 2015;350:117.
2. Dudley J, Deshpande T, Butte A. Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform. 2011;12(4):303–11.
3. Wren J. The emerging in-silico scientist: how text-based bioinformatics is bridging biology and artificial intelligence. IEEE Eng Med Biol Mag. 2004;23(2):87–93.
4. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42:760–772.
5. Altman R. Artificial intelligence (AI) systems for interpreting complex medical data sets. Clin Pharmacol Ther. 2017.
6. Cheng L, Zheng J, Savova G, Erickson B. Discerning tumor status from unstructured MRI reports–completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging. 2010;23(2):119–32.
7. Pham A, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, Morello R, Burgun A. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics. 2014;15:266.
8. Pathak J, Bailey K, Beebe C, Bethard S, Carrell D, Chen P, Dligach D, Endle C, Hart L, Haug P, Huff S, Kaggal V, Li D, Liu H, Marchant K, Masanz J, Miller T, Oniki T, Palmer M, Peterson K, Rea S, Savova G, Stancl C, Sohn S, Solbrig H, Suesse D, Tao C, Taylor D, Westberg L, Wu S, Zhuo N, Chute C. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc. 2013;20(e2):341–8.
9. Lin C, Karlson E, Canhao H, Miller T, Dligach D, Chen P, Perez R, Shen Y, Weinblatt M, Shadick N, Plenge R, Savova G. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS One. 2013;8(8):69932.
10. Ananthakrishnan A, Cai T, Savova G, Cheng S, Chen P, Perez R, Gainer V, Murphy S, Szolovits P, Xia Z, Shaw S, Churchill S, Karlson E, Kohane I, Plenge R, Liao K. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis. 2013;19(7):1411–20.
11. Carroll R, Thompson W, Eyler A, Mandelin A, Cai T, Zink R, Pacheco J, Boomershine C, Lasko T, Xu H, Karlson E, Perez R, Gainer V, Murphy S, Ruderman E, Pope R, Plenge R, Kho A, Liao K, Denny J. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012;19(e1):162–9.
12. Kho A, Hayes M, Rasmussen-Torvik L, Pacheco J, Thompson W, Armstrong L, Denny J, Peissig P, Miller A, Wei W, Bielinski S, Chute C, Leibson C, Jarvik G, Crosslin D, Carlson C, Newton K, Wolf W, Chisholm R, Lowe W. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012;19(2):212–8.
13. Kohane I, Churchill S, Murphy S. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181–5.
14. Wilke R, Xu H, Denny J, Roden D, Krauss R, McCarty C, Davis R, Skaar T, Lamba J, Savova G. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther. 2011;89(3):379–86.
15. McCarty C, Chisholm R, Chute C, Kullo I, Jarvik G, Larson E, Li R, Masys D, Ritchie M, Roden D, Struewing J, Wolf W. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;26(4):13.
16. Waudby C, Berg R, Linneman J, Rasmussen L, Peissig P, Chen L, McCarty C. Cataract research using electronic health records. BMC Ophthalmol. 2011;11:11.
17. Denny J, Ritchie M, Basford M, Pulley J, Bastarache L, Brown-Gentry K, Wang D, Masys D, Roden D, Crawford D. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205–10.
18. Kullo I, Fan J, Pathak J, Savova G, Ali Z, Chute C. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010;17(5):568–74.
19. Liao K, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson E, Plenge R. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010;62(8):1120–7.
20. O'Boyle M. Phelan-McDermid Syndrome Data Network. 2013. http://www.pcori.org/research-results/2013/phelan-mcdermid-syndrome-data-network. [Online; Accessed: February 7,2017].
21. Kohane I, McMurry A, Weber G, MacFadden D, Rappaport L, Kunkel L, Bickel J, Wattanasin N, Spence S, Murphy S, Churchill S. The co-morbidity burden of children and young adults with autism spectrum disorders. PLoS One. 2012;7(4):33224.
22. Collier N. Towards cross-lingual alerting for bursty epidemic events. Journal of Biomedical Semantics. 2011;2(Suppl 5):10.
23. Iris interactive coding system dedicated to the coding of causes of death and to the selection of the underlying causes of death. http://www.cepidc.inserm.fr/site4/index2.php. [Online; Accessed 24 Oct 2017.
24. Névéol A, Dalianis H, Savova G, Zweigenbaum P. Didactic panel: Clinical natural language processing in languages other than English. In: Proc AMIA Annu Symp; 2014.
25. Névéol A, Zweigenbaum P. Clinical natural language processing in 2014: foundational methods supporting efficient healthcare. Yearb Med Inform. 2015;10(1):194–198.
26. Névéol A, Zweigenbaum P. Clinical natural language processing in 2015: Leveraging the variety of texts of clinical interest. Yearb Med Inform. 2016;10(1):234–239.
27. Kreuzthaler M, Schulz S. Detection of sentence boundaries and abbreviations in clinical narratives. BMC Med Inform Decis Mak. 2015;15(Suppl 2):1–13.
28. Oleynik M, Nohama P, Cancian P, Schulz S. Performance analysis of a POS tagger applied to discharge summaries in Portuguese. Stud Health Technol Inform. 2010;160(Pt 2):959–63.
29. Marciniak M. g, Mykowiecka A. Towards morphologically annotated corpus of hospital discharge reports in Polish. In: Proceedings of BioNLP 2011 Workshop. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 92–100. http://www.aclweb.org/anthology/W11-0211.
30. Costumero R, García-Pedrero A, Gonzalo-Martín C, Menasalvas E, Millan S. Text analysis and information extraction from Spanish written documents. In: Slezak D, Tan A.-H, Peters J, Schwabe L, editors. Brain Informatics and Health. Lecture Notes in Computer Science. Springer; 2014. p. 188–197.
31. Baud R, Rassinoux A, Ruch P, Lovis C, Scherrer J. The power and limits of a rule-based morpho-semantic parser. In: Proc AMIA Annu Symp; 1999. p. 22–6.

32. Laippala V, Viljanen T, Airola A, Kanerva J, Salanterä S, Salakoski T, Ginter F. Statistical parsing of varieties of clinical Finnish. Artificial Intelligence In Medicine Special issue: Text Mining and Information Analysis. 2014;61(3):131–6.
33. Nishimoto N, Terae S, Uesugi M, Ogasawara K, Sakurai T. Development of a medical-text parsing algorithm based on character adjacent probability distribution for Japanese radiology reports. Methods Inf Med. 2008;47(6):513–21.
34. Castano J, Gambarte ML, Park HJ, Avila Williams MdP, Perez D, Campos F, Luna D, Benitez S, Berinsky H, Zanetti S. A machine learning approach to clinical terms normalization. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics; 2016. p. 1–11. http://anthology.aclweb.org/W16-2901.
35. Kvist M, Velupillai S. Scan: A swedish clinical abbreviation normalizer. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E, editors. Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science. Springer; 2014. p. 62–73.
36. Kreuzthaler M, Oleynik M, Avian A, Schulz S. Unsupervised Abbreviation Detection in Clinical Narratives. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP). Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p. 91–98. http://aclweb.org/anthology/W16-4213.
37. Shinohara E, Aramaki E, Imai T, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Ohe K. An easily implemented method for abbreviation expansion for the medical domain in Japanese text. a preliminary study. Methods Inf Med. 2013;52(1):51–61.
38. Haverinen K, Ginter F, Viljanen T, Laippala V, Salakoski T. Dependency-based propbanking of clinical finnish. In: Proceedings of the Fourth Linguistic Annotation Workshop. Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 137–141.
39. Palmer M, Kingsbury P, Gildea D. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics. 2005;31:.
40. Haverinen K, Ginter F, Laippala V, Salakoski T. Parsing clinical finnish: Experiments with rule-based and statistical dependency parsers. In: Jokinen K, Bick E, editors. 17th Nordic Conference of Computational Linguistics (NODALIDA 2009). NEALT Proceedings Series. Odense, Denmark; 2009. p. 65–72. Northern European Association for Language Technology.
41. Cohen R, Goldberg Y, Elhadad M. Improving Hebrew segmentation using non-local features with application to information extraction in the medical domain. In: Israeli Seminar on Computational Linguistics; 2010. p. 11–12.
42. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc. 2014;21(5):808–14.
43. Xu Y, Wang Y, Liu T, Liu J, Fan Y, Qian Y, Tsujii J, Chang E. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. J Am Med Inform Assoc. 2014;21(e1):84–92.
44. Cohen R, Goldberg Y, Elhadad M. Transliterated pairs acquisition in medical Hebrew. In: Proc. Machine Translation and Morphologically-rich Languages Workshop; 2011.
45. Engel Thomas C, Bjødstrup Jensen P, Werge T, Brunak S. Negation scope and spelling variation for text-mining of Danish electronic patient records. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi). Gothenburg, Sweden: Association for Computational Linguistics; 2014. p. 64–68.
46. Alicante A, Corazza A, Isgrò F, Silvestri S. Unsupervised information extraction from Italian clinical records. Stud Health Technol Inform. 2007;207:340–9.
47. Alicante A, Corazza A, Isgrò F, Silvestri S. Unsupervised entity and relation extraction from clinical records in italian. Comp Biol Med. 2016;72:263–275.
48. Henriksson A, Moen H, Skeppstedt M, Daudaravičius V, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semantics. 2014;5(1):6.
49. Ahltorp M, Skeppstedt M, Kitajima S, Henriksson A, Rzepka R, Araki K. Expansion of medical vocabularies using distributional semantics on japanese patient blogs. J. Biomedical Semantics. 2016;7:58.
50. Merabti T, Abdoune H, Letord C, Sakji S, Joubert M, Darmoni S. Mapping the ATC classification to the UMLS Metathesaurus: some pragmatic applications. Stud Health Technol Inform. 2011;166:206–13.
51. Perez-de-Viñaspre O, Oronoz M. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. BMC Med Inform Decis Mak. 2015;15(Suppl 2):1–14.
52. Vagelatos A, Mantzari E, Pantazara M, Tsalidis C, Kalamara C. Developing tools and resources for the biomedical domain of the Greek language. Health Informatics J. 2011;17(2):127–39.
53. Hellrich J, Matthies F, Faessler E, Hahn U. Sharing models and tools for processing German clinical texts. In: Stud Health Technol Inform; 2015. p. 734–8.
54. Jimeno Yepes A, Prieur-Gaston E, Névéol A. Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinformatics. 2013;14:146.
55. Neves M, Yepes AJ, Névéol A. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In: Chair NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris, France: European Language Resources Association (ELRA); 2016.
56. Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus developpement. J Biomed Inform. 2014;46(3):506–515.
57. Dalianis H, Velupillai S. De-identifying Swedish clinical text – refinement of a gold standard and experiments with conditional random fields. J Biomed Semantics. 2010;1(1):6.
58. Oronoz M, Gojenola K, Pérez A, de Ilarraza A, A AC. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. J Biomed Inform. 2015;56:318–32.
59. Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French medical corpus: A resource for medical entity recognition and normalization. In: Proc of BioTextMining Workshop, LREC 2014. BioTxtM 2014. Reykjavik, Iceland; 2014. p. 24–30.
60. Grouin C, Lavergne T, Névéol A. Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In: 8th Linguistic Annotation Workshop – LAW VIII, LREC 2014; 2014. p. 54–58.
61. Skeppstedt M, Kvist M, Nilsson G, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. In: Journal of Biomedical Informatics; 2014. p. 148–158.
62. Boytcheva S. Multilingual aspects of information extraction from medical texts in bulgarian. In: Vertan C, von Hahn W, editors. Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation. Cambridge Scholars Publishing; 2012. p. 308–329. http://www.cambridgescholars.com/download/sample/59667.
63. Grigonytė G, Kvist M, Wirèn M, Velupillai S, Henriksson A. Swedification patterns of latin and greek affixes in clinical text. Nordic Journal of Linguistics. 2016;39(01):5–37.
64. Nikolova I, Angelova G. Identifying relations between medical concepts by parsing UMLS definitions. In: Proceedings of the 19th International Conference on Conceptual Structures for Discovering Knowledge. ICCS'11. Berlin, Heidelberg: Springer; 2011. p. 173–186. http://dl.acm.org/citation.cfm?id=2032828.2032843.
65. Deléger L, Merkel M, Zweigenbaum P. Translating medical terminologies through word alignment in parallel text corpora. J Biomed Inform. 2009;42(4):692–701. Epub 2009 Mar 9.
66. Deléger L, Merabti T, Lecrocq T, Joubert M, Zweigenbaum P, Darmoni S. A twofold strategy for translating a medical terminology into French. In: Proc AMIA Annu Symp; 2010. p. 152–6.
67. Drame K, Diallo G, Mougin F. Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus. In: Stud Health Technol Inform; 2012. p. 179–83.
68. Xu Y, Wang Y, Sun J, Zhang J, Tsujii J, Chang E. Building large collections of Chinese and English medical terms from semi-structured and encyclopedia websites. PLoS One. 2013;8(7):67526.
69. Xu Y, Chen L, Wei J, Ananiadou S, Fan Y, Qian Y, Chang E, Tsujii J. Bilingual term alignment from comparable corpora in English discharge summary and Chinese discharge summary. BMC Bioinformatics. 2015;16:149.

70. Kors J, Clematide S, Akhondi S, van Mulligen E, Rebholz-Schuhmann D. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. J Am Med Inform Assoc. 2015;22(5):948–56.
71. Névéol A, Aronson A, Mork J, Darmoni S. Evaluation of French and English MeSH indexing systems with a parallel corpus. In: Proc AMIA Annu Symp; 2005. p. 565–9.
72. Weissenberger C, Jonassen S, Beranek-Chiu J, Neumann M, Müller D, Bartelt S, Schulz S, Mönting J, Henne K, Gitsch G, Witucki G. Breast cancer: patient information needs reflected in English and German web sites. Br J Cancer. 2004;91(8):1482–7.
73. Wu Y, Lei J, Wei W, Tang B, Denny J, Rosenbloom S, Miller R, Giuse D, Zheng K, Xu H. Analyzing differences between Chinese and English clinical text: a cross-institution comparison of discharge summaries in two languages. In: Stud Health Technol Inform; 2013. p. 662–666.
74. Allvin H, Carlsson E, Dalianis H, Danielsson-Ojala R, Daudaravičius V, Hassel M, Kokkinakis D, Lundgrén-Laine H, Nilsson G, Nytrø O, Salanterä S, Skeppstedt M, Suominen H, Velupillai S. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. J Biomed Semantics. 2011;Suppl 3:1.
75. Harris ZS. Language and Information. New York: Columbia University Press; 1988.
76. In: Sager N, Friedman C, Lyman MS, editors. Medical Language Processing: Computer Management of Narrative Data. Reading, MA: Addison Wesley; 1987.
77. Borst F, Sager N, Nhàn NT, Su Y, Lyman M, Tick LJ, Revillard C, Chi E, Scherrer J-R. Analyse automatique de comptes rendus d'hospitalisation. In: Degoulet P, Stéphan J-C, Venot A, Yvon P-J, editors. Informatique et Gestion des Unités de Soins. Informatique et Santé. Springer; 1989. p. 246–256. Chap. 5.
78. Deléger L, Namer F, Zweigenbaum P. Morphosemantic parsing of medical compound words: Transferring a french analyzer to english. International Journal of Medical Informatics. 2009;78 Supplement 1: 48–55. MedInfo 2007.
79. Grouin C, Rosier A, Dameron O, Zweigenbaum P. Testing tactics to localize de-identification. Stud Health Technol Inform. 2009;150:735–9.
80. Becker M, Böckmann B. Extraction of umls®concepts using apache ctakes^{TM} for german language. In: Stud Health Technol Inform; 2016. p. 71–6.
81. Täckström O, McDonald R, J U. Cross-lingual word clusters for direct transfer of linguistic structure. In: Proc NAACL-HLT. Stroudsburg, PA, USA; 2012. p. 477–87.
82. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform. 2009;42(5):923–36.
83. Silva e Oliveira L, de Souza A, Nohama P, Moro C. A rule-based method for continuity of care identification in discharge summaries. In: Stud Health Technol Inform; 2013. p. 1221.
84. Krieger HU, Spurk C, Uszkoreit H, Xu F, Zhang Y, Müller F, Tolxdorff T. Information extraction from German patient records via hybrid parsing and relation extraction strategies. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA); 2014.
85. Moen H, Ginter F, Marsi E, Peltonen L-M, Salakoski T, Salanterä S. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. BMC Med Inform Decis Mak. 2015;15(Suppl 2):1–19.
86. Roque FS, Bjødstrup Jensen P, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, Jensen LJ, Brunak S. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol. 2011;7(8):1002141.
87. Eriksson R, Bjødstrup Jensen P, Frankild S, Juhl Jensen L, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. J Am Med Inform Assoc. 2013;20(5):947–53.
88. Kreuzthaler M, Schulz S, Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. Journal of Biomedical Informatics. 2015;53(Suppl 2):188–195.
89. Gicquel Q, Proux D, Marchal P, Hagége C, Berrouane Y, Darmoni S, Pereira S, Segond F, Metzger M.-H. Évaluation d'un outil d'aide á l'anonymisation des documents médicaux basé sur le traitement automatique du langage naturel. In: Staccini P, Harmel A, Darmoni S, Gouider R, editors. Systèmes D'information Pour L'amélioration de la Qualité en Santé. Informatique et Santé. Springer; 2012. p. 165–176. https://doi.org/10.1007/978-2-8178-0285-5_15. http://dx.doi.org/10.1007/978-2-8178-0285-5_15.
90. Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform. 2014;83(4):303–12.
91. Shin S, Park Y, Shin Y, Choi H, Park J, Lyu Y, Lee M, Choi C, Kim W, Lee J. A de-identification method for bilingual clinical texts of various note types. J Korean Med Sci. 2015;30:7–15.
92. Alfalahi A, Brissman S, Dalianis H. Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In: Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) Held in Conjunction with LREC 2012, May 26, Istanbul; 2012. p. 49–54.
93. Weegar R, Kvist M, Sundström K, Brunak S, Dalianis H. Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx. In: AMIA Annu Symp Proc. San Francisco, USA: AMIA; 2015. p. 1296–305. https://www.ncbi.nlm.nih.gov/pubmed/26958270.
94. Deléger L, Grouin C, Zweigenbaum P. Extracting medication information from French clinical texts. In: Stud Health Technol Inform; 2010. p. 949–953.
95. Boytcheva S. Shallow medication extraction from hospital patient records. In: Stud Health Technol Inform; 2011. p. 119–128.
96. Wang Y, Liu Y, Yu Z, Chen L, Jiang Y. A preliminary work on symptom name recognition from free-text clinical records of traditional chinese medicine using conditional random fields and reasonable features. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. BioNLP '12. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 223–230. http://dl.acm.org/citation.cfm?id=2391123.2391153.
97. Esuli A, Marcheggiani D, Sebastiani F. An enhanced CRFs-based system for information extraction from radiology reports. J Biomed Inform. 2013;46(3):425–35.
98. Segura-Bedmar I, de la Peña González S, Martínez P. Extracting drug indications and adverse drug reactions from Spanish health social media. In: Proceedings of BioNLP 2014. Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 98–106. http://www.aclweb.org/anthology/W/W14/W14-3415.
99. Wang H, Zhang W, Zeng Q, Li Z, Feng K, Liu L. Extracting important information from Chinese operation notes with natural language processing methods. J Biomed Inform. 2014;48:130–6.
100. Nikolova I, Tcharaktchiev D, Boytcheva S, Angelov Z, Angelova G. Applying language technologies on healthcare patient records for better treatment of Bulgarian diabetic patients. In: Agre G, Hitzler P, Krisnadhi A, Kuznetsov S, editors. Artificial Intelligence: Methodology, Systems, and Applications. Lecture Notes in Computer Science. Springer; 2014. p. 92–103.
101. Boytcheva S, Strupchanska A, Paskaleva E, Tcharaktchiev D. Some aspects of negation processing in electronic health records. In: Proceedings of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries; 2005. p. 1–8.
102. Boytcheva S, Tcharaktchiev D, Angelova G. Contextualization in automatic extraction of drugs from hospital patient records. In: Stud Health Technol Inform; 2011. p. 527–31.
103. Zhang S, Kang T, Zhang X, Wen D, Elhadad N, Lei J. Speculation detection for chinese clinical notes: Impacts of word segmentation and embedding models. Journal of Biomedical Informatics. 2016;60:334–341.
104. Skeppstedt M. Negation detection in Swedish clinical text: An adaptation of NegEx to Swedish. J Biomed Semantics. 2011;2(Suppl 3):3.
105. Velupillai S, Skeppstedt M, Kvist M, Mowery D, B C, Dalianis H, Chapman W. Cue-based assertion classification for Swedish clinical text – developing a lexicon for pyConTextSwe, Vol. 61; 2014. p. 137–44.
106. Cotik V, Roller R, Xu F, Uszkoreit H, Budde K, Schmidt D. Negation Detection in Clinical Reports Written in German. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical

Text Mining (BioTxtM2016). Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p. 115–124. http://aclweb.org/anthology/W16-5113.
107. Costumero R, Lopez F, Gonzalo-Martín C, Millan M, Menasalvas E. An Approach to Detect Negation on Medical Documents in Spanish. In: International Conference on Brain Informatics and Health. Springer; 2014. p. 366–375.
108. Cotik V, Stricker V, Vivaldi J, Rodriguez H. Syntactic methods for negation detection in radiology reports in Spanish. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics; 2016. p. 156–165. http://anthology.aclweb.org/W16-2921.
109. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, Conway M, Tharp M, Mowery DL, Deléger L. Extending the negex lexicon for multiple languages. In: MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics, 20-13 August 2013, Copenhagen, Denmark; 2013. p. 677–681.
110. Afzal Z, Pons E, Kang N, Sturkenboom M, Schuemie M, Kors J. Contextd: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. BMC Bioinformatics. 2014;15(1):373.
111. Strötgen J, Gertz M. Multilingual and cross-domain temporal tagging. Language Resources and Evaluation. 2013;47(2):269–298.
112. Tapi Nzali MD, Tannier X, Neveol A. Automatic extraction of time expressions accross domains in french narratives. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics; 2015. p. 492–498. http://aclweb.org/anthology/D15-1055.
113. Velupillai S. Temporal Expressions in Swedish Medical Text – A Pilot Study. In: Proceedings of BioNLP 2014. Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 88–92. http://www.aclweb.org/anthology/W14-3413.
114. Spyns P, Nhan T, Baert E, Sager N, Moor G. Medical language processing applied to extract clinical information from Dutch medical documents. Stud Health Technol Inform. 1998;52(Pt 1):685–689.
115. Hahn U, Romacker M, Schultz S. MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports. Int J Med Inform. 2002;67:63–74.
116. Jonquet C, Musen MA. TOTh'14: Terminology and Ontology: Theories and Applications Workshop. In: Roche C, Costa R, Coudyzer E, editors. Bruxelles, Belgium; 2014. http://hal-lirmm.ccsd.cnrs.fr/lirmm-01099882.
117. Deléger L, Grouin C, Ligozat A-L, Zweigenbaum P, Névéol A. Annotation of specialized corpora using a comprehensive entity and relation scheme. In: Proc of LREC; 2014. p. 1267–1274.
118. Bretschneider C, Zillner S, Hammon M. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. Sofia, Bulgaria: Association for Computational Linguistics; 2013. p. 27–35. http://www.aclweb.org/anthology/W13-1904.
119. Figueroa R, Soto D, Pino E. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. In: Conf Proc IEEE Eng Med Biol Soc; 2014. p. 2710–3.
120. Grouin C, Deléger L, Rosier A, Temal L, Dameron O, Van Hille P, Burgun A, Zweigenbaum P. Automatic computation of CHA2DS2-VASc score: information extraction from clinical texts for thromboembolism risk assessment. In: Proc AMIA Annu Symp; 2011. p. 501–10.
121. Kontio E, Airola A, Pahikkala T, Lundgren-Laine H, Junttila K, Korvenranta H, Salakoski T, Salanterä S. Predicting patient acuity from electronic patient records. J Biomed Inform. 2014;51:35–40.
122. Hu D, Huang Z, Chan T-M, Dong W, Lu X, Duan H. Utilizing chinese admission records for mace prediction of acute coronary syndrome. International Journal of Environmental Research and Public Health. 2016;13(9):912.
123. Jacobson O, Dalianis H. Applying deep learning on electronic health records in swedish to predict healthcare-associated infections. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics; 2016. p. 191–195. http://anthology.aclweb.org/W16-2926.
124. Löpprich M, Krauss F, Ganzinger M, Senghas K, Riezler S, Knaup P. Automated Classification of Selected Data Elements from Free-text Diagnostic Reports for Clinical Research. Method Inf Med. 2016;55(4):373–80.
125. Khachidze M, Tsintsadze M, Archuadze M. Natural language processing based instrument for classification of free text medical records. Biomed Res Int. 20168313454.
126. Metzger M, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a french pilot study. Int J Methods Psychiatr Res. 2016.
127. Takano K, Ueno M, Moriya J, Mori M, Nishiguchi Y, Raes F. Unraveling the linguistic nature of specific autobiographical memories using a computerized classification algorithm. Behavior Research Methods. 20161–18.
128. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah N. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. J Am Med Inform Assoc. 2015;22(1):121–31.
129. Smith L, Rindflesch T, Wilbur W. The importance of the lexicon in tagging biological text. Natural Language Engineering. 2005;12(2):1–17.
130. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Computational Biology. 2013;9(2):1002854.
131. McClosky D, Charniak E. Self-training for biomedical parsing. In: Proceedings of ACL-08: HLT, Short Papers. Columbus, Ohio: Association for Computational Linguistics; 2008. p. 101–104. http://www.aclweb.org/anthology/P08-2026.
132. Hara T, Miyao Y, Tsujii J. Evaluating impact of re-training a lexical disambiguation model on domain adaptation of an HPSG parser. In: Proceedings of IWPT. Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 11–22.
133. Aronson A, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36.
134. Savova G, Masanz J, Ogren P, Zheng J, Sohn S, Kipper-Schuler K, Chute C. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
135. cTAKES natural language processing system for extraction of information from electronic medical record clinical free-text. http://www.ctakes.apache.org. [Online; Accessed 24 Oct 2017].
136. Jonquet C, Shah N, Musen M. The open biomedical annotator. In: Summit on Translat Bioinforma; 2009. p. 56–60.
137. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):267–70.
138. Rebholz-Schuhmann D, Clematide S, Rinaldi F, Kafkas S, van Mulligen E, Bui C, Hellrich J, Lewin I, Milward D, Poprat M, Jimeno-Yepes A, Hahn U, Kors J. Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In: Forner P, Müller H, Paredes R, Rosso P, Stein B, editors. Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Lecture Notes in Computer Science. Springer; 2013. p. 353–367.
139. Cornet R. A Dutch treat for healthcare terminology. In: Proc CLEF 2013 Evaluation Labs and Workshop – CLEF-ER 2013; 2013.
140. Névéol A, Grosjean J, Darmoni S, Zweigenbaum P. Language resources for French in the biomedical domain. In: Proc Language and Resource Evaluation Conference, LREC 2014; 2014. p. 2146–2151.
141. Schulz S, Ingenerf J, Thun S, Daumke P. German-language content in biomedical vocabularies. In: Proc CLEF 2013 Evaluation Labs and Workshop – CLEF-ER 2013; 2013.
142. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007;14(5):550–63.
143. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17(5):514–8.
144. Elhadad N, Pradhan S, Gorman S, Manandhar S, Chapman W, Savova G. SemEval-2015 task 14: Analysis of clinical text. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics; 2015. p. 303–310. http://www.aclweb.org/anthology/S15-2051.
145. Styler WF IV, Bethard S, Finan S, Palmer M, Pradhan S, de Groen P, Erickson B, Miller T, Lin C, Savova G, Pustejovsky J. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist. 2014;2:143–54.
146. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 task 6: Clinical TempEval. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics; 2015. p. 806–814. http://www.aclweb.org/anthology/S15-2136.

147. Saeed M, Villaroel M, Reisner A. T, Clifford G, Lehman L.-W, Moody G, Heldt T, Kyaw T. H, Moody B, Mark R. G. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Crit Care Med. 2011;39(5):952–60.
148. Turku Clinical TreeBank and PropBank. http://bionlp.utu.fi/clinicalcorpus.html. [Online; Accessed 24 Oct 2017].
149. Aramaki E, Morita M, Kano Y, Ohkuma T. Overview of the NTCIR-12 MedNLPDoc task. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo Japan; 2016.
150. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011;18(5):540–543.
151. Aramaki E, Morita M, Kano Y, Ohkuma T. Overview of the NTCIR-11 MedNLP-2 task. In: Proceedings of the 11th NTCIR Conference. Tokyo Japan; 2014.
152. Névéol A, Grouin C, Tannier X, Hamon T, Kelly L, Goeuriot L, Zweigenbaum P. CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: CLEF 2015 Online Working Notes. CEUR-WS; 2015.
153. Névéol A, Cohen K. B, Grouin C, Hamon T, Lavergne T, Kelly L, Goeuriot L, Rey G, Robert A, Tannier X, Zweigenbaum P. Clinical information extraction at the CLEF eHealth Evaluation lab 2016. In: CLEF 2016 Online Working Notes. CEUR-WS; 2016. p. 28–42.
154. Balk E, Chung M, Chen M, Chang L, Trikalinos T. Data extraction from machine-translated versus original language randomized trial reports: a comparative study. Syst Rev. 20132–97.
155. Zeng-Treitler Q, Kim H, Rosemblat G, Keselman A. Can multilingual machine translation help make medical record content more comprehensible to patients?. Stud Health Technol Inform. 2010;160 (Pt 1):73–77.
156. Wu C, Xia F, Deléger L, Solti I. Statistical machine translation for biomedical text: are we there yet?. In: Proc AMIA Annu Symp; 2011. p. 1290–9.
157. Liu W, Cai S. Translating electronic health record notes from English to Spanish: A preliminary study. In: Proceedings of BioNLP 15. Beijing, China: Association for Computational Linguistics; 2015. p. 134–140. http://www.aclweb.org/anthology/W15-3816.
158. Rayner E, Bouillon P, Brotanek J, Flores G, Halimi Mallem IS, Hockey BA, Isahara H, Kanzaki K, Kron E, Nakao Y, Santaholma ME, Starlander M, Tsourakis N. The MEDSLT 2008 system. In: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications, COLING 2008; 2008. p. 32–35.
159. Névéol A, Max A, Ivanishcheva Y, Ravaud P, Zweigenbaum P, Yvon F. Statistical machine translation of systematic reviews into French. In: Proc Workshop on Optimizing Understanding in Multilingual Hospital Encounters – TIA 2013; 2013. p. 10–13.
160. Thirion B, Pereira S, Névéol A, Dahamna B, Darmoni S. French MeSH browser: a cross-language tool to access MEDLINE/PubMed. In: Proc AMIA Annu Symp; 2007. p. 1132.
161. Fontelo P, Liu F, Leon S, Anne A, Ackerman M. PICO Linguist and BabelMeSH: development and partial evaluation of evidence-based multilanguage search tools for MEDLINE/PubMed. Stud Health Technol Inform. 2007;129(Pt 1):817–821.
162. Mahmoud M, Al-Khafaji J, Al-Shorbaji N, Sara K, Al-Ubaydli M, Ghazzaoui R, Liu F, Fontelo P. BabelMeSH and PICO Linguist in Arabic. In: AMIA Annu Symp Proc; 2008. p. 944.
163. Pecina P, Dušek O, Goeuriot L, Hajič J, Hlaváčová J, Jones G, Kelly L, L eveling J, Mareček D, Novák M, Popel M, Rosa R, Tamchyna A, Urešová Z. Adaptation of machine translation for multilingual information retrieval in the medical domain. Artif Intell Med. 2014;61(3):165–85.
164. Camacho-Collados J, Pilehvar MT, Navigli R. A unified multilingual semantic representation of concepts. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015). Beijing, China: Association for Computational Linguistics; 2015. p. 741–751.
165. Duque A, Martinez-Romo J, Araujo L. Can multilinguality improve biomedical word sense disambiguation?. Journal of Biomedical Informatics. 2016;64:320–332.
166. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, Clark C. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One. 2014;9(11):112774.
167. Low Resource Languages for Emergent Incidents. http://www.darpa.mil/program/low-resource-languages-for-emergent-incident%s. [Online; Accessed 24 Oct 2017].